# Useful Features for Human Verification in Near-Infrared Periocular Images

Karen Hollingsworth, Kevin W. Bowyer, and Patrick J. Flynn

*University of Notre Dame, Notre Dame, IN 46556*

**Abstract**

The periocular region is the part of the face immediately surrounding the eye, and researchers have recently begun to investigate how to use the periocular region for recognition. Understanding how humans recognize faces helped computer vision researchers develop algorithms for face recognition. Likewise, understanding how humans analyze periocular images could benefit researchers developing algorithms for periocular recognition. We conducted two experiments to determine how humans analyze periocular images. In these experiments, we presented pairs of images and asked volunteers to determine whether the two images showed eyes from the same subject or from different subjects. In the first experiment, subjects were paired randomly to create different-subject queries. Our volunteers correctly determined the relationship between the two images in 92% of the queries. In the second experiment, we considered multiple factors in forming different-subject pairs; queries were formed from pairs of subjects with the same gender and race, and with similar eye color, makeup, eyelash length, and eye occlusion. In addition, we limited the amount of time volunteers could view a query pair. On this harder experiment, the correct verification rate was 79%. We asked volunteers to describe what features in the images were helpful to them in making their decisions. In both experiments, eyelashes were reported to be the most helpful feature.

*Keywords:* periocular recognition, biometrics, near-infrared light

## 1. Introduction

The periocular region is the part of the face immediately surrounding the eye. While the face and the iris have both been studied extensively as biometric characteristics [1, 2], the use of the periocular region for a biometric system is an emerging field of research. Periocular biometrics could potentially be combined with iris biometrics to obtain a more robust system than iris biometrics alone. If an iris biometrics system captured an iris image of poor quality, the region surrounding the eye might still be used to confirm or refute an identity. A further argument for researching periocular biometrics is that current iris biometric systems already capture images containing some periocular information, yet when making recognition decisions, they ignore all pixel information outside the iris region. The periocular area of the image may contain useful information that could improve recognition performance, if we could identify and extract useful features in that region.

A few papers [3, 4, 5, 6, 7, 8, 9, 10] have presented algorithms for periocular recognition, but their approaches have relied on general computer vision techniques rather than methods specific to this biometric characteristic. One way to begin designing algorithms specific to this region of the face is to examine how humans make recognition decisions using the periocular region.

Other computational vision problems have benefitted from a good understanding of the human visual system. In a recent book chapter, O'Toole [11] says, "Collaborative interactions between computational and psychological approaches to face recognition have offered numerous insights into the kinds of face representations capable of supporting the many tasks humans accomplish with faces" [11]. Sinha et al. [12] describe numerous basic findings from the study of human face recognition that have direct implications for the design of computational systems. Their report says "The only system that [works] well in the face of [challenges like sensor noise, viewing distance, and illumination] is the human visual system. It makes eminent sense, therefore, to attempt to understand the strategies this biological system employs, as a first step towards eventually translating them into machine-based algorithms" [12].

In this study, we investigated which features humans found useful for making decisions about identity based on periocular information. We presented pairs of images

Table I: Research in Image Classification

| Paper | Data | Algorithm | Features |
|---|---|---|---|
| Abiantun and Savvides [13] | ICE data: 2953 near-infrared iris images | Classify images as right or left eyes. Extract features using adaboost. Classify using support vector machines, linear discriminant analysis, or principal component analysis. | tear-duct |
| Bhat and Savvides [14] | ICE data: 2953 near-infrared iris images; additional near-infrared iris images | Classify images as right or left eyes using active shape models. | eye shape |
| Li et al. [15] | CMU-PIER data: 107 East-Asian subjects; subset of UBIRISv1 data: 107 Caucasian subjects | Classify images as Asian or Caucasian using active shape models, edge filters, and a nearest neighbor classifier. | eyelashes |
| Merkow et al. [16] | Images downloaded from the web | Gender classification using local binary patterns (LBP), principal component analysis, linear discriminant analysis, and support vector machines. | LBP features |
| Lyle et al. [17] | FRGC data: visible light face images: 410 subjects | Gender and ethnicity classification using grayscale pixel intensities and local binary patterns with a non-linear support vector machine. | pixel intensity, LBP features |

to volunteers and asked them to determine whether the two images showed eyes from the same subject or from different subjects. In our first experiment, subjects were paired randomly to create different-subject queries. In the second experiment, we challenged participants by pairing similar images together rather than pairing subjects at random for the different-subject queries. In both experiments, we asked volunteers to describe what features in the images were helpful to them in making their decisions. We found that the features that humans found most helpful were not the features used by current periocular biometrics work [3, 4, 5, 6, 7, 8, 9, 10]. Based on our research, we anticipate that explicit modeling and description of eyelids, eyelashes, and tear ducts could yield more recognition power than the current periocular biometrics algorithms published in the literature.

The rest of this paper is organized as follows. Section 2 summarizes the previous work in periocular biometrics. Section 3 describes how we selected and pre-processed eye images for our experiment. Our experimental method is outlined in Section 4. Section 5 presents our analysis. Finally, Section 6 presents a summary of our findings, a discussion of the implications of our experiment, and recommendations for future work.

## 2. Related Work

The work related to periocular biometrics can be classified into two categories. The first category includes research in segmenting and describing periocular features for image classification. This research classifies images as containing left or right eyes, or it classifies images by gender or ethnicity. Works in this category are listed in Table I.

The second category includes research that has analyzed periocular features for recognition purposes. These works used gradient orientation histograms, local binary patterns, and SIFT features for periocular recognition. Works in this category are listed in Table II.

Table II: Research in Periocular Recognition

| Paper | Data | Algorithm | Features |
|---|---|---|---|
| Park et al. [3] | 899 visible light face images, 30 subjects | Gradient orientation histograms, local binary patterns, Euclidean distance, SIFT matcher | Eye region with width of 6*iris-radius and height of 4*iris-radius |
| Miller et al. [4] | FRGC data: visible light face images, 410 subjects; FERET data: visible light face images, 54 subjects | Local binary patterns, City block distance | LBP features |
| Adams et al. [5] | Same as Miller et al. | Local binary patterns and genetic algorithm to select features | LBP features |
| Woodard et al. [7] | FRGC data: visible light face images; MBGC near-infrared face images | Local binary patterns with city block distance and color features with Bhattacharya coefficient | LBP features, color |
| Woodard et al. [6] | MBGC data: near infrared face images, 88 subjects | Local binary patterns; Result fused with iris matching results | LBP features |
| Miller et al. [8] | FRGC data: visible light face images | Local binary patterns, color | LBP features, color |
| Xu et al. [9] | FRGC data: visible light face images | Local Walsh-Transform binary patterns | LBP features |
| Bharadwaj et al. [10] | UBIRISv2: visible light iris images | Circular local binary patterns and second-order global statistics | LBP features and second-order features |
| This work | Near infrared images from LG 2200 iris camera | Human analysis | Eyelashes, tear duct, eyelids |

One difference between our work and the above mentioned papers is the target data type. The periocular recognition papers all used periocular regions cropped from face data. Our work uses near infrared images of a small periocular region, from the type of image we get from iris cameras. The anticipated application is to use periocular information to assist in iris recognition when iris quality is poor.

Another difference between our work and the above work is the development strategy. The recognition papers have followed a strategy of applying common computer vision techniques to analyze images. We attempted to approach periocular recognition from a different angle. We aimed to investigate the features that humans find most useful for recognition in near infrared images of the periocular region.

## 3. Data

In selecting our data, we considered using eye images taken from two different cameras: an LG2200 and an LG4000 iris camera. The LG2200 is an older model, and the images taken with this camera sometimes have undesirable interlacing or lighting artifacts [18]. On the other hand, in our data sets, the LG4000 images seemed to show less periocular data around the eyes. Since our purpose was to investigate features in the periocular region, we chose to use the LG2200 images so that the view of the periocular region would be larger. We hand-selected a subset of images, choosing images in good focus, with minimal interlacing and shadow artifacts. We also favored images that included both the inner and outer corners of the eye.

For our first experiment, we selected images from 120 different subjects. We had 60 male subjects and 60 female subjects. 108 of them were Caucasian and 12 were Asian.

For 40 of the subjects, we selected two images of an eye and saved the images as a "match" pair. In each case, the two images selected were acquired at least a week apart. For the remaining subjects, we selected one image of an eye, paired it with an image from another subject, and saved it as a "nonmatch" pair. Thus, the queries that we would present to our volunteers involved 40 match pairs, and 40 nonmatch pairs. All queries were either both left eyes, or both right eyes.

In our second experiment, we used images from 210 subjects. We had 104 male and 106 female subjects. 187 were Caucasian, 15 Asian, 3 Asian-Southern, 3 Hispanic, and 2 Black or African-American. As in our previous experiment, we randomly assigned subjects to be used in either "match" or "nonmatch" pairs. We had 70 subjects for the "match" pairs and 140 subjects for the "nonmatch" pairs. Rather than randomly pairing the 140 nonmatch subjects into queries, we paired similar subjects together. All nonmatch subjects were paired so that two subjects in a pair had the same gender and race. In addition, similar subjects were paired as follows. For all possible pairs of images, we computed a difference score based on eye color (blue, green, hazel, light brown, or dark brown), presence of makeup (no-makeup, light-makeup, or heavy-makeup), dilation ratio, percent eye occlusion, eyelashes (short, medium, or long), and contacts (present or absent). We then paired the most-similar subjects together to make nonmatch queries. For match queries, we used images taken at least a week apart so that no query would show images from the same session. Additionally, we randomly chose whether to show two left eyes or two right eyes for the query.

In both experiments, our objective was to examine how humans analyzed the periocular region. Consequently, we did not want the iris to be visible during our tests. To locate the iris in each image, we used our automatic segmentation software, which uses active contours to find the iris boundaries. Next, we hand-checked all of the segmentations. If our software had made an error in finding the inner or outer iris boundary, we manually marked the center and a point on the boundary to identify the correct center and radius of an appropriate circle. If the software had made an error in finding the eyelid, we marked four points along the boundary to define three line segments approximating the eyelid contour.

For all of the images, we set the pixels inside the iris/pupil region to black. An example image where the iris has been blacked-out is shown in Figure 6.

## 4. Experimental Method

In order to determine which features in the periocular region were most helpful to the human visual system, we designed an experiment to present pairs of eye images to volunteers and ask for responses. We designed a graphical user interface (GUI) to display our images. At the beginning of each session, the computer displayed example pairs of eye images to the participant. The examples included both match and nonmatch pairs. Next, the computer displayed the test queries. For each test query, the software displayed a pair of images and asked the user to respond whether he or she thought the two images were from the same person or from different people. In addition, he could note his level of confidence in his response – whether he was "certain" of his response, or only thought that his response was "likely" the correct answer.

The user was asked to rate a number of features depending on whether each feature was "very helpful," "helpful," or "not helpful" for determining identity. The features listed were "eye shape," "tear duct,"[1] "outer corner," "eyelashes," "skin," "eyebrow," "eyelid," and "other." If a user marked that some "other" feature was helpful, he was asked to enter what feature(s) he was referring to. A final text box on the screen asked the user to describe any other additional information that he used while examining the eye images.

In Experiment 1, we asked volunteers to rate the helpfulness of the various features for every single query pair. Users did not have any time limit for examining the images. After the user had classified the pair of images as "same person" or "different people" and rated all features, he could click "Next" to proceed. At that point the user was told whether he had correctly classified the pair of images. Then, the next query was displayed. All volunteers saw the same queries, but the order of the queries was randomized for each volunteer. A screenshot of the GUI interface is visible in Figure 1.

One drawback of our first experimental design was that the number of queries was relatively small. Despite the small number of queries, we had one participant take an hour and forty minutes to respond to the 80 queries. In order to present a larger number of queries during Experiment 2, we limited viewing time to three seconds for each pair of images. By limiting viewing times, we could show a larger number of queries and therefore get feedback

[1] We used the term "tear duct" informally in this instance to refer to the region near the inner corner of the eye. A more appropriate term might be "medial canthus" but we did not expect the volunteers in our experiment to know this term.

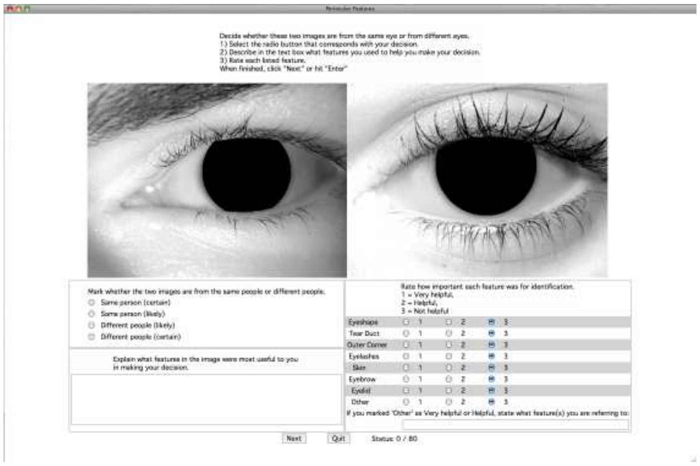

Figure 1: In our first experiment, we asked participants to rate the helpfulness of various features with every single query.

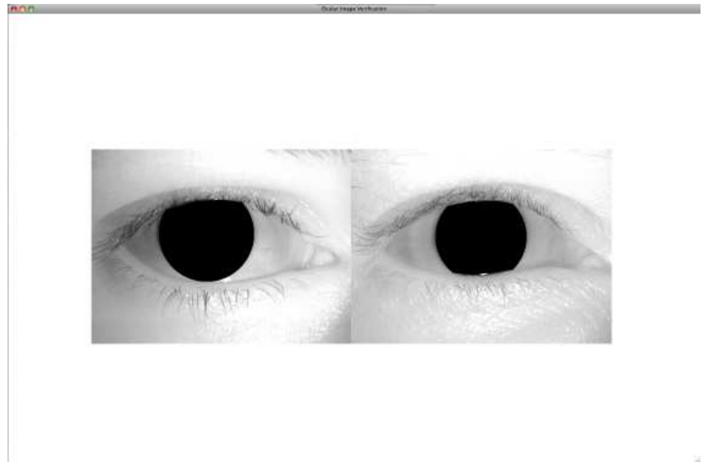

Figure 2: In our second experiment, the software displayed a pair of images like the pair shown above for three seconds. After the three seconds, the images were hidden and users could respond whether the two images were same or different. We asked participants to rate the helpfulness of various features only once at the end of the experiment, after they had seen all of the queries.

on valuable features after participants had seem a greater number of images. Unfortunately, this design makes it difficult to directly compare accuracy on Experiments 1 and 2; however, we are able to achieve our primary goal of determining which features humans find most useful. A reasonable area of future work would be to determine how much of the accuracy difference between the two experiments is due to the shortened viewing time and how much is due to pairing similar images together for non-match queries.

Experiment 2 showed pairs of images for three seconds, and after the allotted time, the images were hidden from view. At that point, users could respond whether the two images were from the same person or from different people. Once the user responded and clicked "Next", the software reported whether the user had correctly classified the pair of images. As in Experiment 1, the order of the queries was randomized for each user. Users were only asked to rate the helpfulness of the various features once, after seeing all of the queries. A screenshot of the GUI interface is visible in Figure 2.

We solicited volunteers from the students and staff at the University of Notre Dame to participate in our experiments. We had 25 volunteers participate in Experiment 1, and 28 volunteers for Experiment 2.

## 5. Results

### *5.1. How well can humans determine whether two periocular images are from the same person or not?*

To find overall accuracy scores for our experiments, we counted the number of times the participant was "likely" or "certain" of the correct response; that is, we made no distinction based on the participant's confidence level, only on whether they believed a pair to be from the same person, or believed a pair to be from different people. We divided the number of correct responses by the total number of queries to yield and accuracy score.

On Experiment 1, the average number of correct responses was 73.68 out of 80, which is 92.10% (standard deviation 4.62%). The minimum score was $\frac{65}{80} = 81.25\%$, and the maximum score was $\frac{79}{80} = 98.75\%$.

On Experiment 2, the average number of correct responses was 110.25 out of 140, which is 78.75% (standard deviation 5.70%). The minimum score was $\frac{89}{140} = 63.57\%$, and the maximum score was $\frac{124}{140} = 88.57\%$.

We used a t-test to evaluate the null hypothesis that humans did not perform differently than random guessing. For both experiments, the resulting p-value was less than $10^{-4}$. Thus, we have statistically significant evidence that our volunteers were doing better than random.

### *5.2. Did humans score higher when they felt more certain?*

As mentioned above, users had the option to mark whether they were "certain" of their response or whether their response was merely "likely" to be correct. Some participants were more "certain" than others.

On Experiment 1, one participant responded "certain" for 70 of the 80 queries. On the other hand, one participant did not answer "certain" for any queries. Discounting the person who was never certain, the average score on the questions where participants were certain was 96.64% (standard deviation 5.26%). The average score when participants were less certain was 84.84% (standard deviation 11.24%).

On Experiment 2, one participant responded "certain" for 140 of the 140 queries. Two responded "certain" for only 19 of the 140 queries. The average score on the questions where participants were certain was 88.90% (standard deviation 7.46%). The average score when participants were less certain was 70.77% (standard deviation 8.09%). Thus, in both experiments volunteers did better on the subset of the queries where they felt "certain" of their answer.

### *5.3. Did humans do better on the second half of the test than the first half?*

To determine whether participants were improving throughout the duration of the experiment, we compared scores from the first half of the test with scores from the second half of the test. For Experiment 1, the average scores on the two halves of the test were nearly identical. On the first half of queries the average score was 92.20% (standard deviation 5.12%), and on the second half of the queries, the average score was 92.00% (standard deviation 5.00%). For Experiment 2 where we presented more queries, there was some improvement between the two halves of the test. On the first half of queries the average score was 77.86% (standard deviation 5.30%), and on the second half of the queries, the average score was 79.64% (standard deviation 7.85%). We computed a one-tailed t-test to check whether the scores on the second half were statistically signicantly higher than the scores on the first half. The resulting p-value was 0.095. Thus there is insufficient evidence to show that the subjects learned over the course of the test.

These results are consistent with results from other experiments where we had participants view periocular regions of twins' eyes [19] and of left and right eye pairs [20]. In both cases, the average scores were higher on the second half of the test, but we did not find statistically significant evidence of improvement. It may be that a longer test is needed in order to see statistically significant evidence of learning.

### *5.4. In Experiment 1, which features were correlated with correct responses?*

A primary goal of our research was to determine which features in the periocular region were most helpful to the human visual system when making recognition decisions. Specifically, we are interested in features present in near-infrared images of the type that can be obtained by a typical iris camera. In Experiment 1, we asked participants to rate the helpfulness of features on every query; therefore, we could evaluate which features they reported as useful on the subset of queries where they *correctly* determined whether the image pair was from same person.

For all correct responses, we counted the number of times each feature was rated as "very helpful" to the user, "helpful", or "not helpful". A bar chart of these counts is given in Figure 3. The features in this figure are sorted by the number of times each feature was regarded as "very helpful". According to these results, the most helpful feature was eyelashes, although tear duct and eye shape were also very helpful. The ranking from most helpful to least helpful was (1) eyelashes, (2) tear duct, (3) eye shape, (4) eyelid, (5) eyebrow, (6) outer corner, (7) skin, and (8) other.

Other researchers have found eyebrows to be more useful than eyes in identifying famous people [12], so the fact that eyebrows were ranked fifth out of eight is perhaps deceiving. The reason eyebrows received such a low ranking in our experiment is that none of the images showed a complete eyebrow. In about forty queries, the two images both showed some part of the eyebrow, but in the other forty queries, the eyebrow was outside the image field-of-view in at least one of the images in the pair. On images with a larger field of view, eyebrows could be significantly more valuable. We suggest that iris sensors with a larger field of view would be more useful when attempting to combine iris and periocular biometric information.

The low ranking for "outer corner" (sixth out of eight) did not surprise us, because in our own observation of a number of eye images, the outer corner does not often provide much unique detail for distinguishing one eye from another. There were three queries where the outer corner of the eye was not visible in the image (See Figure 9).

Skin ranked seventh out of eight in our experiment, followed only by "other". Part of the reason for the low rank of this feature is that the images were all near-infrared images. Therefore, participants could not use skin color to make their decisions. This result may not be quite as striking if we used a data set containing a greater diversity of ethnicities. However, we have noticed that variations in lighting can make light skin appear dark in a near-infrared image, suggesting that overall intensity in the skin region may have greater intra-class variation than inter-class variation in these types of images.

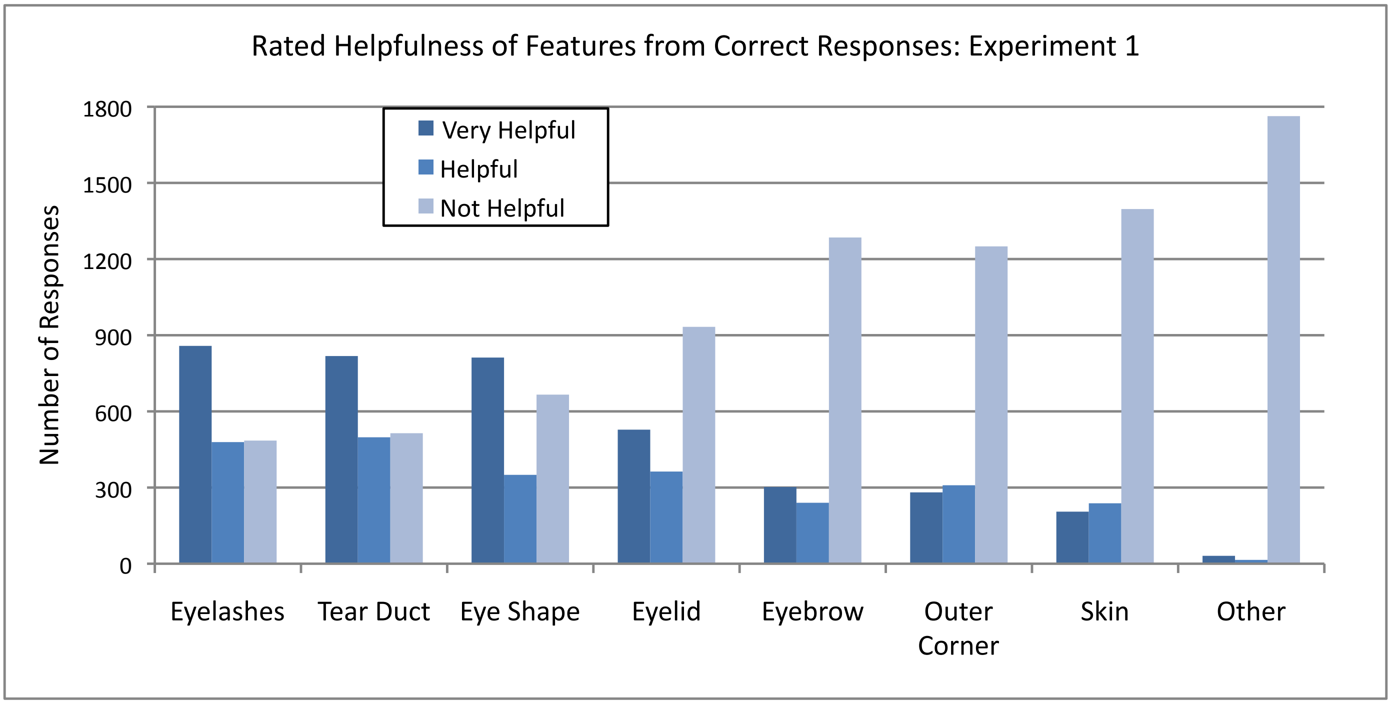


Figure 3: Eyelashes were considered the most helpful feature for making decisions about identity. The tear duct and shape of the eye were also very helpful.

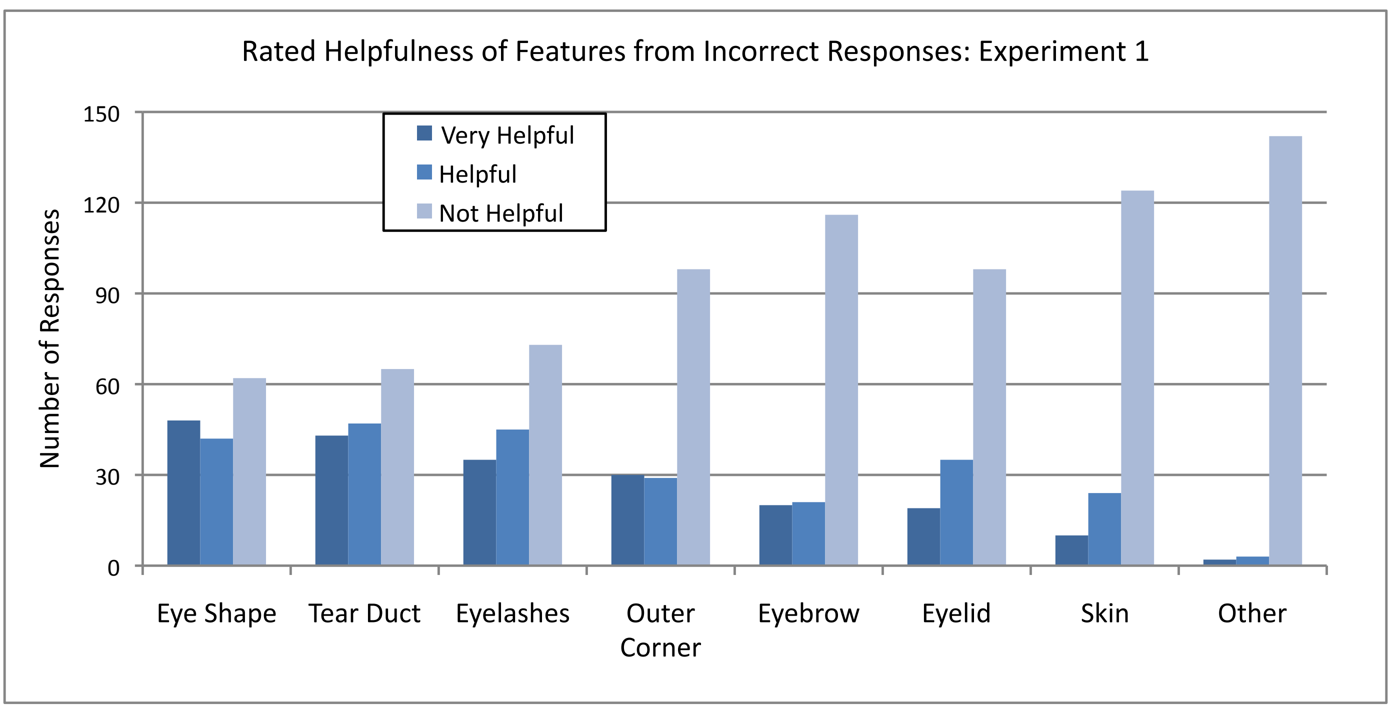


Figure 4: We compared the rankings for the features from correct responses (Fig. 3) with the rankings from incorrect responses. The shape of the eye and the outer corner of the eye were both used more frequently on incorrect responses than on correct responses. This result suggests that those two features would be less helpful for making decisions about identity than other features such as eyelashes.

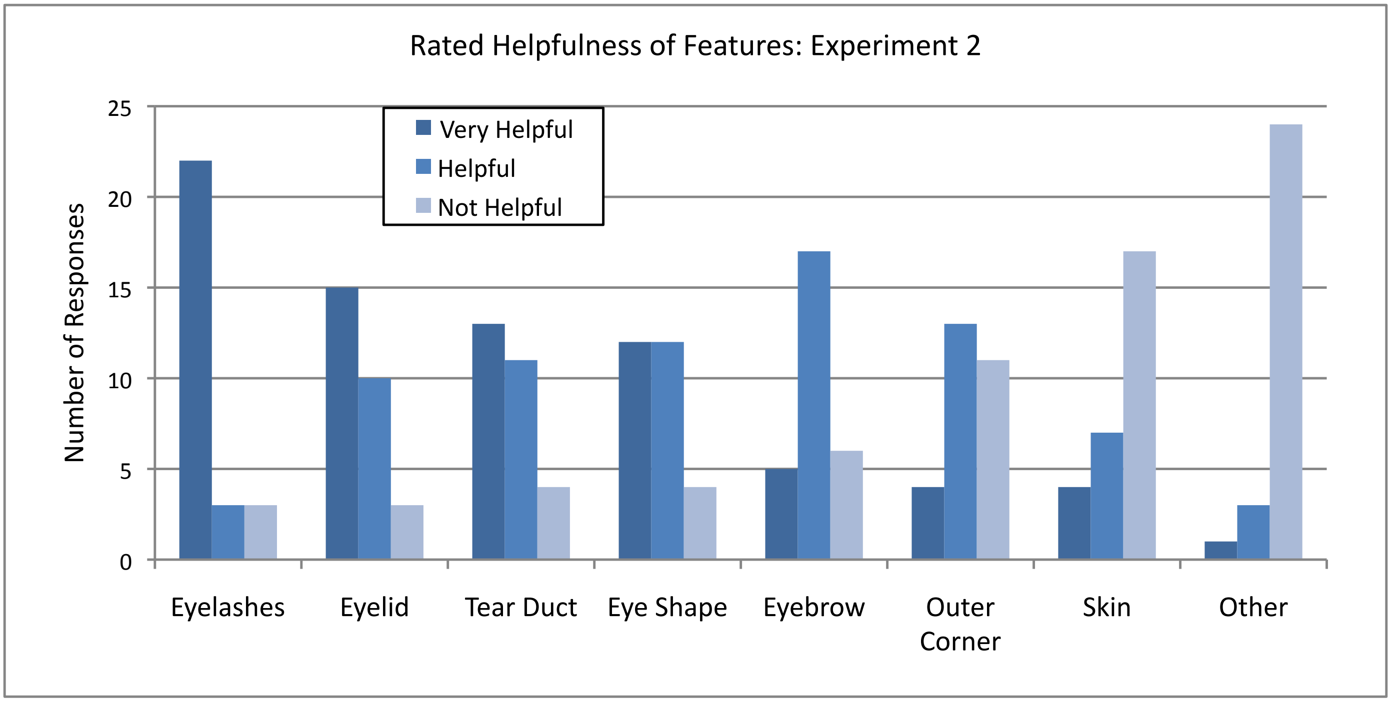


Figure 5: In both Experiment 1 (Fig. 3) and Experiment 2, eyelashes were the top-ranked feature. The tear duct was also very valuable, being ranked second in Experiment 1 and third in Experiment 2.

### 5.5. *In Experiment 1, which features were correlated with incorrect responses?*

In addition to considering which features were marked most helpful for correct responses, we also looked at how features were rated when participants responded incorrectly. For all the *incorrectly* answered queries, we counted the number of times each feature was "very helpful", "helpful", or "not helpful". A bar chart of these counts is given in Figure 4. We might expect to have a similar rank ordering for the features in the incorrect queries as we had for the correct queries, simply because if certain features are working well for identification, a volunteer would tend to continue to use the same features. Therefore, rather than focusing on the overall rank order of the features, we considered how the feature rankings *differed* from the correct responses to the incorrect responses. The ranking from most helpful feature to least helpful feature for the incorrect queries was (1) eye shape, (2) tear duct, (3) eyelashes, (4) outer corner, (5) eyebrow, (6) eyelid, (7) skin, and (8) other. Notice that "eye shape" changed from rank three to rank one. Also "outer corner" changed from rank six to rank four. This result implies that eye shape and outer corner are features that are less valuable for correct identification. On the other hand, "eyelashes" and "eyelid" both changed rank in the opposite direction, implying that those features are more valuable for correct identification.

### 5.6. *In Experiment 2, which features were rated most helpful?*

In Experiment 2, we asked participants to rate features after they had seen all of the queries. Thus, we could not analyze which features they found useful on specific queries. However, we did tabulate the feature rankings that participants gave. A bar chart of these counts is given in Figure 5. In both Experiment 1 and Experiment 2, eyelashes were rated as the most helpful feature. The next three features – tear duct, eye shape, and eyelid – were ranked in slightly different order for the two experiments, but still ranked in the top four. The consistency of responses over two different experiments with different images and different participants shows that eyelashes are very helpful to humans in verification tasks with these types of images. The tear duct is also very valuable, being ranked second in Experiment 1 and third in Experiment 2.

### 5.7. *What additional information did humans provide?*

In addition to the specific features that participants were asked to rate, participants were also asked to describe other factors they considered in making their decisions. Users were prompted to "explain what features in the image were most useful to you in making your decision", and enter their response in a text box.

Table III: Summary of Responses
to an Open-Ended Request to list Most Useful Features

| Query Type | Helpful Features | Unhelpful or Misleading Features |
|---|---|---|
| Match Queries | clusters of eyelashes | glare |
| | single "stray" eyelashes | shadow |
| | eyelash density | different lighting |
| | eyelash direction | different angle of eye |
| | eyelash length | different eye shape |
| | eyelash intensity | amount the eye was open |
| | tear duct | hair in one image |
| | eyebrow | contact lens |
| | unusual eye shape | vs. no contact lens |
| | slant of eyes | makeup vs. no makeup |
| | amount the eye was open | |
| | contacts | |
| | makeup | |
| Nonmatch Queries | lashes in tear duct region | glare |
| | eyelash density | makeup |
| | eyelash direction | |
| | eyelash length | |
| | eyelash intensity | |
| | tear duct | |
| | eyebrow | |
| | eyelid | |
| | eye shape | |
| | crease above the eye | |
| | contacts | |
| | makeup | |

Table III summarizes volunteers' free-responses. Only responses from queries where they got the answer correct are listed. Participants found a number of different traits of eyelashes valuable. They considered the density of eyelashes (or number of eyelashes), eyelash direction, length, and intensity (light vs. dark). Clusters of eyelashes, or single eyelashes pointing in an unusual direction were helpful, too. Contacts were helpful as a "soft biometric". That is, the presence of a contact lens in both images could be used as supporting evidence that the two images were of the same eye. However, no participants relied on contacts as a deciding factor. Two of the eighty queries in Experiment 1 showed match pairs where one image in the pair showed a contact lens, and the other did not. Participants did well for both of these pairs: the percents of volunteers who classified these pairs correctly were 92% (23 of 25) and 96% (24 of 25).

Makeup was listed both as "very helpful" for some queries, and as "misleading" for other queries. When a subject wore exactly the same type of makeup for multiple acquisition sessions, the makeup was useful for recognition. Alternatively, when a subject changed her makeup, recognition was harder. One of the eighty queries in Experiment 1 showed a match pair where only one of the images displayed makeup. Although 24 of 25 participants still correctly classified this pair, every participant who provided written comments for this pair remarked that the presence of mascara in only one of the images was distracting or misleading.

### 5.8. *Which pairs were most frequently classified correctly, and which pairs were most frequently classified incorrectly?*

In Experiment 1, there were 21 match pairs that were classified correctly by all participants. One example of a pair that was classified correctly by all participants is shown in Figure 6. There were 12 nonmatch pairs classified correctly by all participants. An example is shown in Figure 7.

Figure 8 shows the match pair most frequently classified incorrectly in Experiment 1. Eleven of the 25 participants mistakenly thought that these two images were from different people. This pair is challenging because the eye is wide open in one of the images, but not it the other. Figure 9 shows the nonmatch pair most frequently classified incorrectly. This pair was also misclassified by 11 participants, although the set of 11 participants who responded incorrectly for the pair in Figure 9 was different from the set of participants who responded incorrectly for Figure 8.

In Experiment 2, there were 3 match pairs that were classified correctly by all volunteers. One example of a pair that was classified correctly by all volunteers is shown in Figure 10. There were no nonmatch pairs classified correctly by all volunteers, but Figure 11 shows a nonmatch pair classified correctly by 27 of 28 volunteers.

Figure 12 shows the match pair most frequently classified incorrectly in Experiment 2. Seventeen of the 25 volunteers mistakenly thought that these two images were from different people. Figure 13 shows the nonmatch pair most frequently classified incorrectly. This pair was misclassified by 16 volunteers.

## 6. Discussion and Conclusion

We conducted two experiments examining how well humans could classify a pair of periocular images as being from the same person or from different people. In Experiment 1, we formed nonmatch queries by randomly pairing two subjects together. In Experiment 2, we formed nonmatch queries by pairing subjects with the same gender, same ethnicity, and similar eye color, makeup, eye occlusion, and eyelash length. Also in Experiment 2, we limited the viewing time to three seconds for each pair to allow us to present a larger number of queries in the experiment. In both experiments, we presented an equal number of match and nonmatch queries. We found that humans correctly classified the pairs on the easier task (Experiment 1) with an average accuracy of 92%. On the harder task (Experiment 2), average accuracy was 79%. Thus, we observed a large drop in performance on the harder task. However, both experiments showed humans performing significantly better than random guessing.

Participants' scores were higher on the queries where they expressed high confidence. On the subset of queries where participants were confident, the average score was 97% for Experiment 1 and 89% for Experiment 2. Therefore, we infer that participants correctly judged their relative confidence in their responses.

The performance on Experiment 1 was about 92% for both the first and second halves of the test. However, on Experiment 2, which presented 75% more queries, performance improved by about 2% between the first and second portions of the test. This improvement was not statistically significant, but it is possible that a longer test might show statistically significant evidence of learning.

Eyelashes were rated as the most helpful feature in both Experiments 1 and 2. Participants used eyelash intensity, length, direction, and density. They also looked for groups of eyelashes that clustered together, and for single eyelashes separated from the others. The tear duct was rated as the second most helpful feature in Experiment 1, and the third most helpful feature in Experiment 2. Eye shape and eyelids were also rated highly. However, eye shape was used in a large number of incorrect responses. Both eye shape and the outer corner of the eye were used a higher proportion of the time for incorrect responses than they were for correct responses, thus those two features might not be as useful for recognition. Skin and the outer corner of the eye were ranked lowest in both experiments.

The presence of contacts was used as a soft biometric. Eye makeup was helpful in some image pairs, and distracting in others. Changes in lighting were challenging, and large differences in eye occlusion were also a challenge.

Our analysis suggests some specific ways to design powerful periocular biometrics systems. We expect that for near-infrared periocular images, a biometrics system that explicitly detects eyelids, eyelashes, the tear duct and the entire shape of the eye could be more powerful than some of the skin analysis methods presented previously.

While the eyelashes were judged the most helpful feature, analyzing the eyelashes would likely require detecting the eyelids first. Eyelids can be detected using edge detection and Hough transforms [21, 22], a parabolic "integrodifferential operator" [23], or active contours [24]. The research into eyelid detection has primarily been aimed at detecting and disregarding the eyelids during iris recognition, but we suggest detecting and describing eyelids and eyelashes to aid in identification. Feature vectors describing eyelashes could include measures for the density of eyelashes along the eyelid, the uniformity of direction of the eyelashes, and the curvature and length of the eyelashes. We could also use metrics comparing the upper and lower lashes.

The second most helpful feature in our study was the

Figure 6: In Experiment 1, all 25 participants correctly classified these two images as being from the same person.

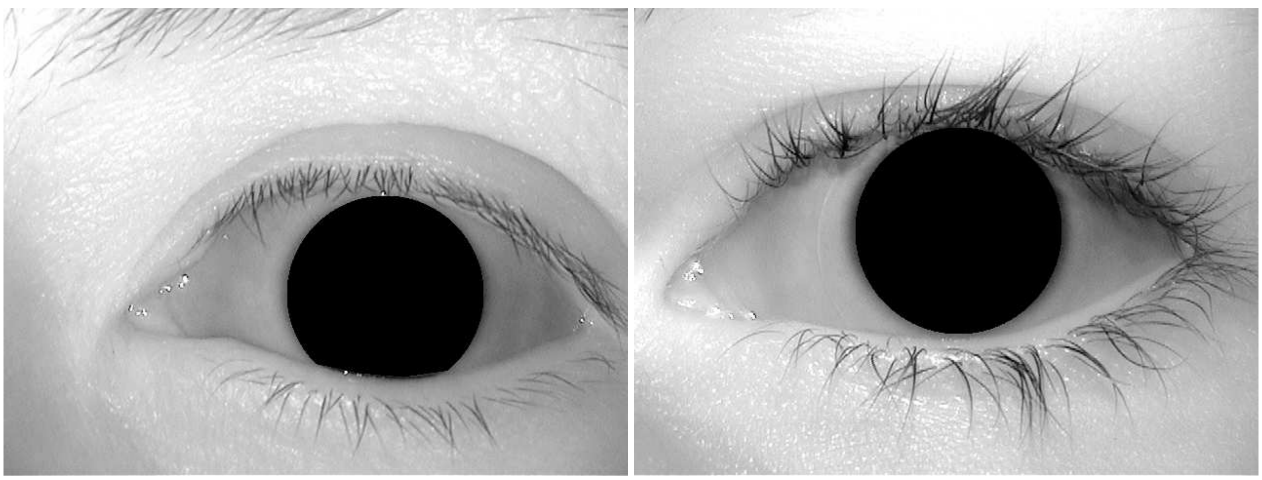

Figure 7: In Experiment 1, all 25 participants correctly classified these two images as being from different people

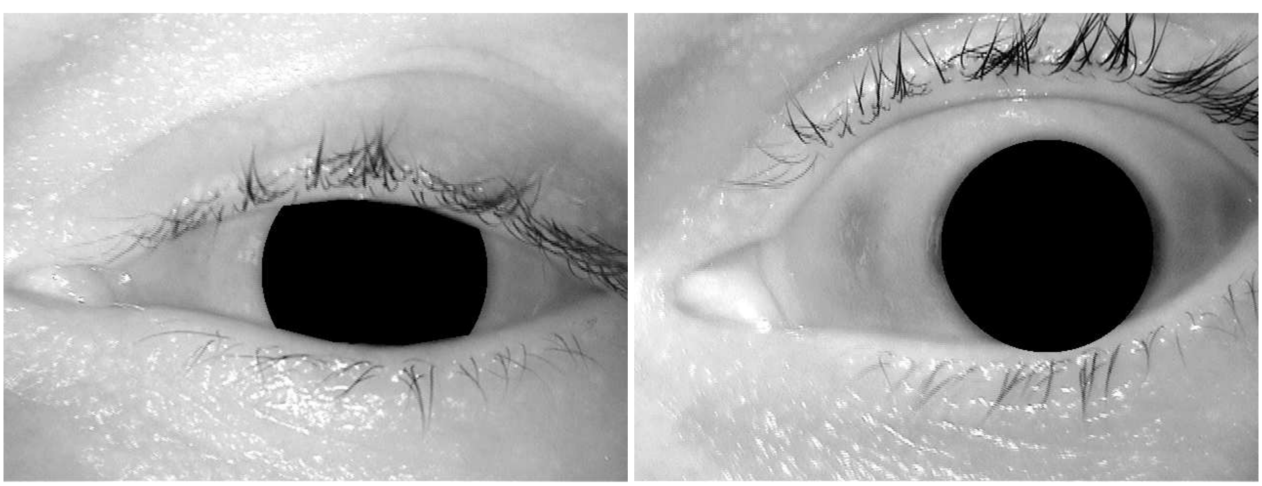

Figure 8: In Experiment 1, eleven of 25 participants incorrectly guessed that these images were from different people, when in fact, these eyes are from the same person. This pair is challenging because one eye is much more open than the other.

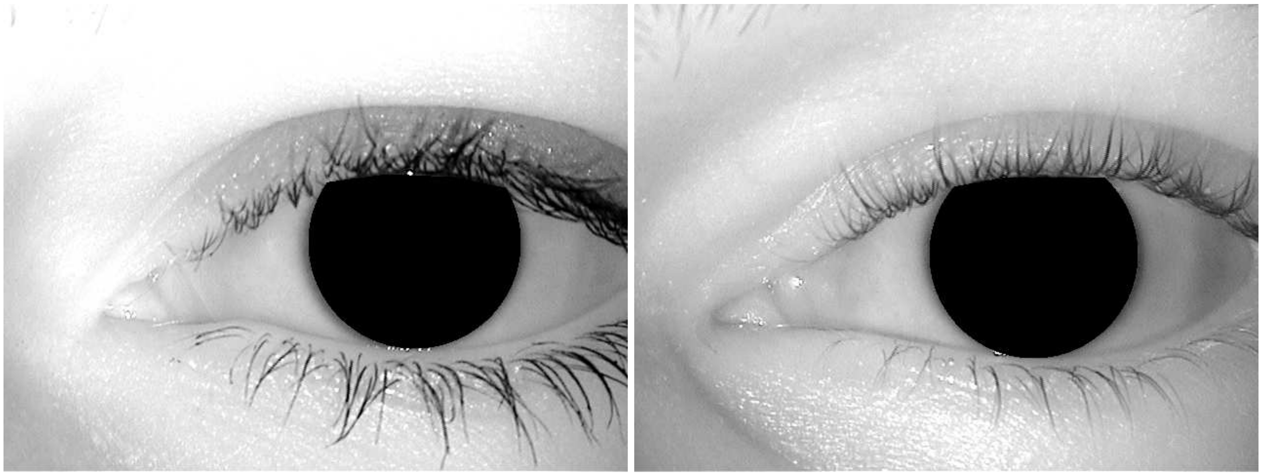

Figure 9: In Experiment 1, eleven of 25 participants incorrectly guessed that these images were from the same person, when in fact, they are from two different people.

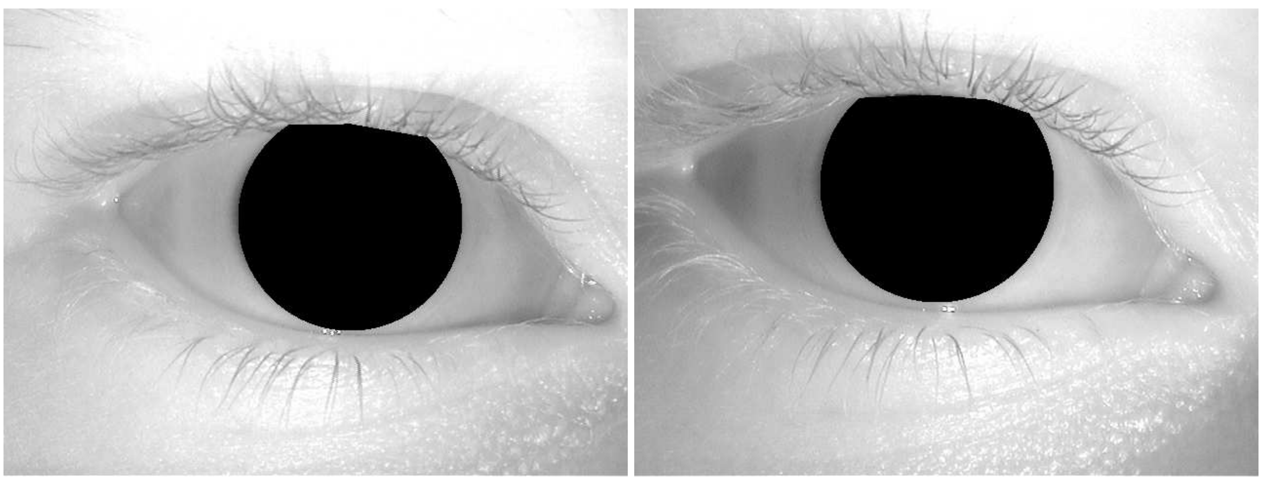

Figure 10: All 28 participants in Experiment 2 correctly classified these two images as being from the same person.

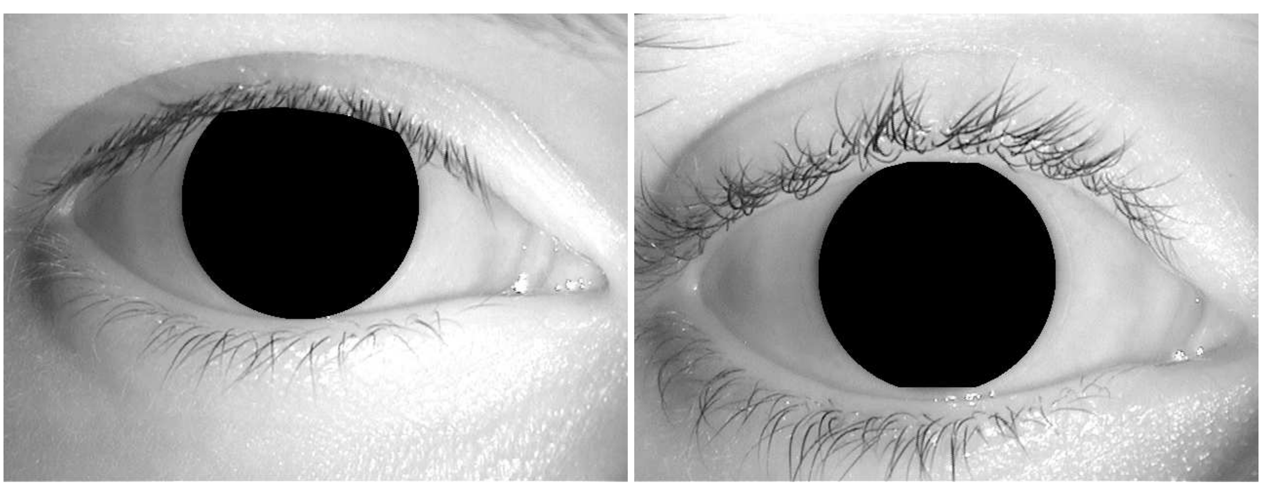

Figure 11: Twenty-seven of 28 participants in Experiment 2 correctly classified these two images as being from different people

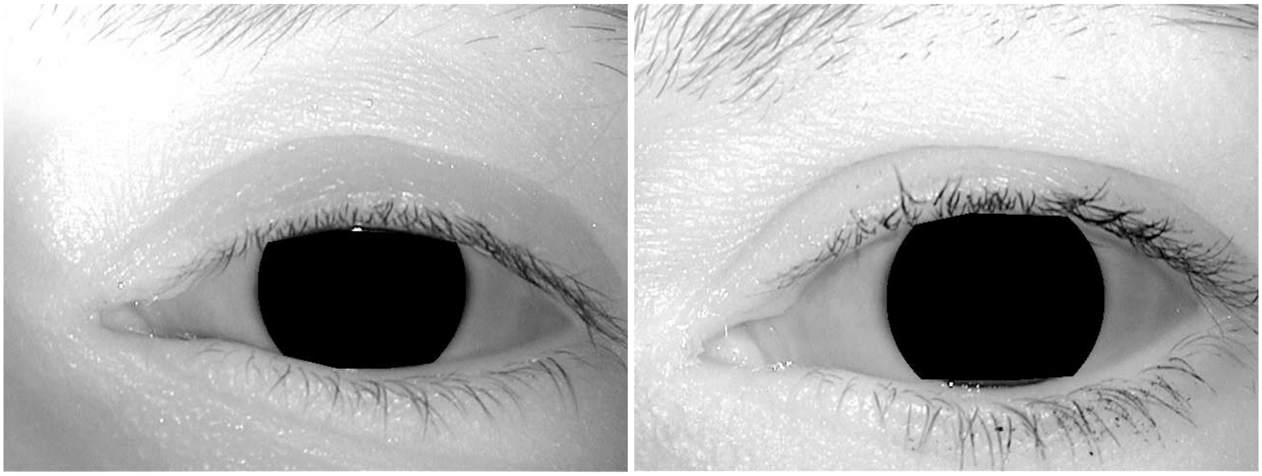

Figure 12: Seventeen out of 28 participants in Experiment 2 incorrectly guessed that these images were from different people, when in fact, these eyes are from the same person.

Figure 13: Sixteen of 28 participants in Experiment 2 incorrectly guessed that these images were from the same person, when in fact, they are from two different people.

tear duct region. Once we have detected the eyelids, we could extend those curves to locate the tear duct region. This region should more formally be referred to as the *medial canthus*. A canthus is the angle or corner on each side of the eye, where the upper and lower lids meet. The medial canthus is the inner corner of the eye, or the corner closest to the nose. Two structures are often visible in the medial canthus, the *lacrimal caruncle* and the *plica semilunaris* [25]. These two features typically have lower contrast than eyelashes and iris. Therefore, they would be harder for a computer vision algorithm to identify, but if they were detectable, the sizes and shapes of these structures would be possible features. Detecting the medial canthus itself would be easier than detecting the caruncle and plica semilunaris, because the algorithm could follow the curves of the upper and lower eyelids until they meet at the canthus. Alternatively, we could follow the method suggested by Abiantun and Savvides [13] using boosted Haar features. Once detected, we could measure the angle formed by the upper and lower eyelids and analyze how the canthus meets the eyelids. In Asians, the epicanthal fold may cover part of the medial canthus [25] so that there is a smooth line from the upper eyelid to the inner corner of the eye (e.g. Figure 6). The epicanthal fold is present in fetuses of all races, but in Caucasians it has usually disappeared by the time of birth [25]. Therefore, Caucasian eyes are more likely to have a distinct cusp where the medial canthus and upper eyelid meet (e.g. Figure 8).

The shape of the eye has potential to be helpful, but the term "eye shape" is ambiguous, which might explain the seemingly contradictory results we obtained about the helpfulness of this particular feature. To describe the shape of the eye, we could analyze the curvature of the eyelids. We could also detect the presence or absence of the *superior palpebral furrow* – the crease in the upper eyelid – and measure its curvature if present.

Previous periocular research has focused on texture and key points in the area around the eye. The majority of prior work [4, 5, 6, 7, 8] masked an elliptical region in the middle of the periocular region "to eliminate the effect of textures in the iris and the surrounding sclera area" [4]. This mask effectively occludes a large portion of the eyelashes and tear duct region, thus hiding the features that we find are most valuable. Park et al. [3] do not mask the eye, but they also do not do any explicit feature modeling beyond detecting the iris. These promising prior works have all shown recognition rates at or above 77%. However, we suggest that there is potential for greater recognition power by considering additional features.

## 7. Acknowledgements

This work is supported by the Federal Bureau of Investigation, the Central Intelligence Agency, the Intelligence Advanced Research Projects Activity, the Biometrics Task Force, the Technical Support Working Group through US Army contract W91CRB-08-C-0093, the Intelligence Community Postdoctoral Fellowship Program and by the CIA award US-2010-1048708-000. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of our sponsors.

## References

[1] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys 35 (4) (2003) 399–458.

[2] K. W. Bowyer, K. P. Hollingsworth, P. J. Flynn, Image understanding for iris biometrics: A survey, Computer Vision and Image Understanding 110 (2) (2008) 281–307.

[3] U. Park, A. Ross, A. K. Jain, Periocular biometrics in the visible spectrum: A feasibility study, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2009, pp. 1–6.

[4] P. Miller, A. Rawls, S. Pundlik, D. Woodard, Personal identification using periocular skin texture, in: Proc. ACM 25th Symposium on Applied Computing (SAC2010), 2010, pp. 1496–1500.

[5] J. Adams, D. L. Woodard, G. Dozier, P. Miller, K. Bryant, G. Glenn, Genetic-based type II feature extraction for periocular biometric recognition: Less is more, in: Proc. Int. Conf. on Pattern Recognition, 2010, pp. 205–208.

[6] D. L. Woodard, S. Pundlik, P. Miller, R. Jillela, A. Ross, On the fusion of periocular and iris biometrics in non-ideal imagery, in: Proc. Int. Conf. on Pattern Recognition, 2010, pp. 201–204.

[7] D. L. Woodard, S. J. Pundlik, J. R. Lyle, P. E. Miller, Periocular region appearance cues for biometric identification, in: IEEE Computer Vision and Pattern Recognition Biometrics Workshop, 2010, pp. 162–169.

[8] P. E. Miller, J. R. Lyle, S. J. Pundlik, D. L. Woodard, Performance evaluation of local appearance based periocular recognition, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2010, pp. 1–6.

[9] J. Xu, M. Cha, J. L. Heyman, S. Venugopalan, R. Abiantun, M. Savvides, Robust local binary pattern feature sets for periocular biometric identification, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2010, pp. 1–8.

[10] S. Bharadwaj, H. S. Bhatt, M. Vatsa, R. Singh, Periocular biometrics: When iris recognition fails, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2010, pp. 1–6.

[11] A. Calder, G. Rhodes (Eds.), Handbook of Face Perception, Oxford University Press, 2010, Ch. Cognitive and Computational Approaches to Face Perception by O'Toole, in press.

[12] P. Sinha, B. Balas, Y. Ostrovsky, R. Russell, Face recognition by humans: Nineteen results all computer vision researchers should know about, Proceedings of the IEEE 94 (11) (2006) 1948–1962.

[13] R. Abiantun, M. Savvides, Tear-duct detector for identifying left versus right iris images, in: 37th IEEE Applied Imagery Pattern Recognition Workshop (AIPR2008), 2008, pp. 1–4.

[14] S. Bhat, M. Savvides, Evaluating active shape models for eye-shape classification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2008), 2008, pp. 5228–5231.

[15] Y. hui Li, M. Savvides, T. Chen, Investigating useful and distinguishing features around the eyelash region, in: Proceedings of the 2008 37th IEEE Applied Imagery Pattern Recognition Workshop, 2008, pp. 1–6.

[16] J. Merkow, B. Jou, M. Savvides, An exploration of gender identification using only the periocular region, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2010, pp. 1–5.

[17] J. R. Lyle, P. E. Miller, S. J. Pundlik, D. L. Woodard, Soft biometric classification using periocular region features, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2010, pp. 1–7.

[18] K. W. Bowyer, P. J. Flynn, The ND-IRIS-0405 iris image dataset, Tech. rep., University of Notre Dame, http://www.nd.edu/~cvrl/papers/ND-IRIS-0405.pdf (2009).

[19] K. Hollingsworth, K. W. Bowyer, P. J. Flynn, Similarity of iris texture between identical twins, in: IEEE Computer Vision and Pattern Recognition Biometrics Workshop, 2010, pp. 1–8.

[20] K. W. Bowyer, S. Lagree, S. Fenker, Human versus biometric detection of similarity in left and right irises, in: IEEE International Carnahan Conference on Security Technology (ICCST), 2010.

[21] R. P. Wildes, Iris recognition: An emerging biometric technology, Proceedings of the IEEE 85 (9) (1997) 1348–1363.

[22] B. J. Kang, K. R. Park, A robust eyelash detection based on iris focus assessment, Pattern Recognition Letters 28 (13) (2007) 1630–1639. doi:10.1016/j.patrec.2007.04.004.

[23] J. Daugman, How iris recognition works, IEEE Transactions on Circuits and Systems for Video Technology 14 (1) (2004) 21–30.

[24] W. J. Ryan, D. L. Woodard, A. T. Duchowski, S. T. Birchfield, Adapting starburst for elliptical iris segmentation, in: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems, 2008, pp. 1–7.

[25] C. Oyster, The Human Eye Structure and Function, Sinauer Associates, 1999.